

# Sistemas de arquivos distribuídos e replicados em rede com alta disponibilidade em ambiente Open Source



**Darlan Segalin**

# Agenda

- **Ambiente**
- **Planejamento**
- **DRBD**
- **Open Source Cluster File System**
- **How does it work?**
- **Alta Disponibilidade**
- **Considerações Finais**

# Ambiente

**É o ambiente formado por algumas tecnologias:**

- **Servidores Linux**
- **Protocolos Comunicação**
- **Replicação Dados**
- **Alta Disponibilidade**
- **Monitoramento**

# Duvida!

**É de grátis?**

(Seu Creysson)

# Planejamento

- **Onde Usar?**
- **Quando?**
- **Requisitos.**

# What is DRBD?



- **É um dispositivo de blocos designado para construir clusters altamente disponíveis.**
- **Network Raid 1.**
- **Necessário aplicativos que funcionem em cima de um dispositivo de blocos DRBD.**
- **Exemplos: um File System & Fsck, um Journaling FS.**

# DRBD (How does it work ?)



## DRBD versão 0.7.x.

- Cada dispositivo tem um estado.
- Primeiro nó como dispositivo primário é montado por `/dev/drbdX`.
- Toda gravação local também é enviado para o nó secundário.

# DRBD (How does it work ?)



## DRBD versão 0.7.x.

- **Caso o primeiro nó falhar, é necessário mudar o dispositivo secundário em primário.**
- **Quando o nó com falha voltar a funcionar, os dados serão sincronizados para ele e o mesmo ficará no modo secundário.**
- **Sincronização Inteligente versão 0.7 séries, até 4TB.**

# DRBD (How does it work ?)



- **DRBD versão 0.7.x.**
- **Clusters HA (HP, IBM, ...) usam SCSI buses or Fibre Channel para dispositivos de blocos compartilhados.**
- **DRBD usa mesma semântica rodando em cima de redes IP.**
- **Recomendados usar sistema de arquivos com Journaling (Ex: Ext3, JFS, XFS).**

# DRBD (How does it work ?)



## DRBD versão 8.x.

- **Todos nodos pode ser montados com o papel principal (primário).**
- **Ativo x Ativo.**
- **Sistemas de arquivos paralelo (Cluster File System), OCFS OpenGFS ou GFS.**

# DRBDLinks

- **Programa onde é definidos links de diretórios que apontarão para o dispositivos de bloco compartilhados.**

Ex. arquivo de configuração:

```
mountpoint('/shared')
```

```
link('/home/')
```

```
link('/etc/httpd')
```

# What is OCFS2?

**General purpose cluster file system.**

- **Modo de disco compartilhado.**
- **Utilizado em conjunto com DRBD.**
- **Nós replicados e primários (podem ser montados como RWX em todos nós de cluster).**
- **Com base no POSIX compliant file system.**

# Why use OCFS2?


- **TCP-based.**
- **Melhorou Journaling.**
- **Melhorou performance (Space Allocation).**
- **Melhorou cache de dados.**
- **Network based pluggable DLM.**
- **OCFS2 requer Kernel 2.6.x**


# ocfs2console


Node Configuration


Nodes:


Active	Name	Node	IP Address	IP Port
	rac1		10.1.10.191	7777
	rac2		10.1.10.192	7777

 Add

 Edit

 Remove

 Apply

 Close

# OCFS2 and DRBD uses

- **Servidor de arquivos.**
- **FTP.**
- **NFS.**
- **HTTP.**
- **Mysql.**
- **Oracle Databases.**
- **Xen Image Migration.**
- **...**

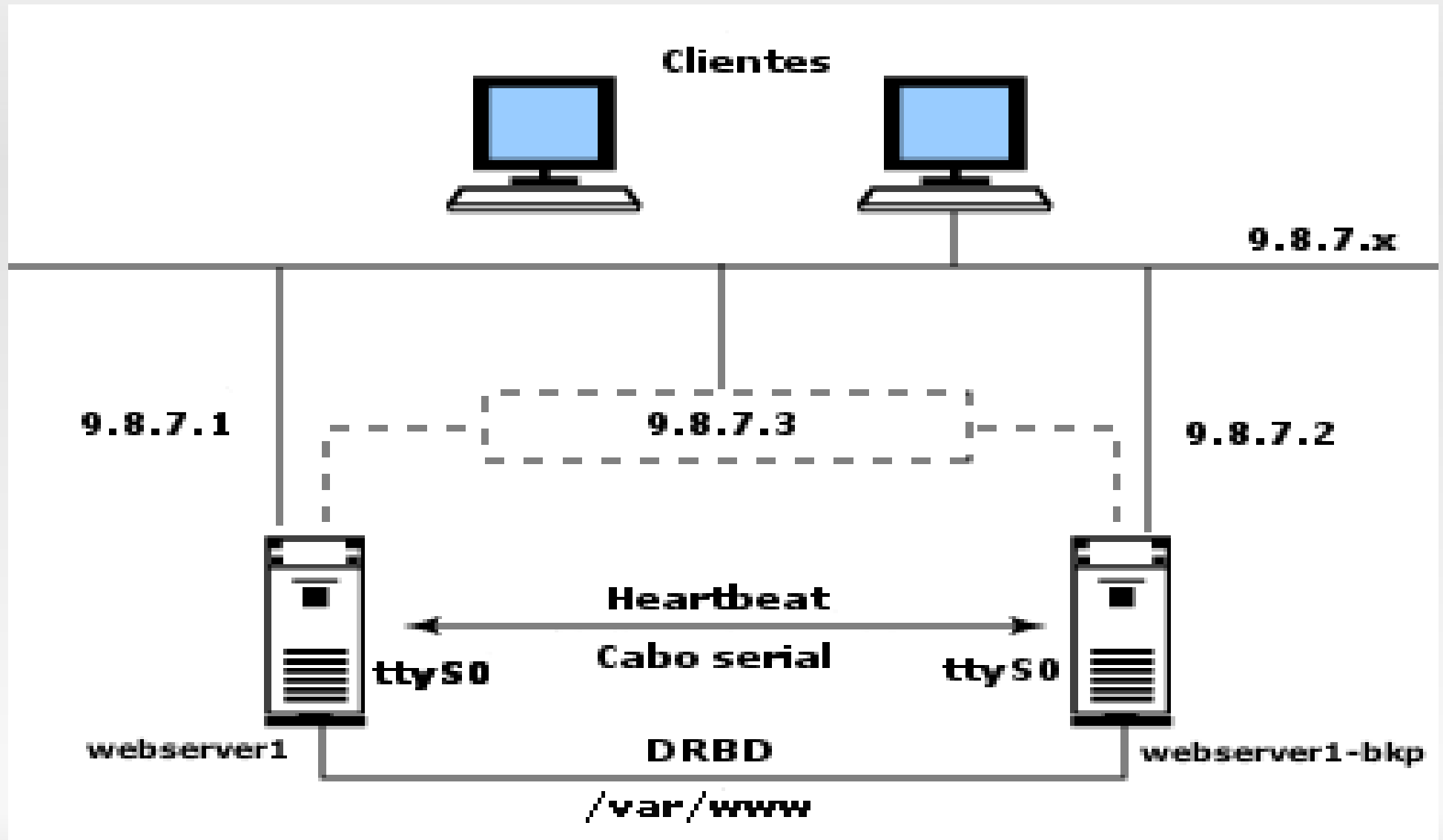
# Heartbeat

- **Heartbeat significa batimento cardíaco.**
- **Envia sinais através de serial, ethernet ou ambas, se o heartbeat falhar, a máquina secundária irá assumir que a primária falhou, e tomar os serviços que estavam rodando na máquina primária.**



- **Heartbeat controla a inicialização de determinados serviços e recursos dos servidores.**
- **Não iniciar serviços controlados pelo heartbeat automaticamente no boot do linux.**
- **<http://www.linux-ha.org>**

# Heartbeat



# Tabela Disponibilidade

Disponibilidade (%)	<i>Downtime/ano</i>	<i>Downtime/mês</i>
95%	18 dias 6:00:00	1 dias 12:00:00
96%	14 dias 14:24:00	1 dias 4:48:00
97%	10 dias 22:48:00	0 dias 21:36:00
98%	7 dias 7:12:00	0 dias 14:24:00
99%	3 dias 15:36:00	0 dias 7:12:00
99,9%	0 dias 8:45:35.99	0 dias 0:43:11.99
99,99%	0 dias 0:52:33.60	0 dias 0:04:19.20
99,999%	0 dias 0:05:15.36	0 dias 0:00:25.92

# Monitoramento (Mon)

- **Ferramenta para monitorar o estado do cluster e enviar avisos via email, celular, pager, etc.**
- **Pode ser implementado via software script SHELL.**

# Talk is cheap...



**...show me the implementation.**

# Requisitos

- **Kernel Source ou headers.**
- **2 Servidores - máquinas idênticas se possível.**
- **Rede Gibabit.**
- **128 MB de disco em cada nó será usado para metadados DRBD.**
- **Obtendo DRBD:**
- **<http://oss.linbit.com/drbd>**

# Instalação DRBD



**Descompacta pacote**

```
# cp drbd-8.0pre3.tar.gz /usr/src
```

```
# cd /usr/src
```

```
# tar xzvf drbd-8.0pre3.tar.gz
```

**Indica onde está código fonte.**

```
# cd drbd08/drbd
```

```
# make KDIR=/usr/src/código-fonte-  
do-kernel
```

```
# make install
```

# Instalação DRBD



**Instala e compila as ferramentas administrativas (drbdsetup, drbdadm, drbdmeta)**

```
# cd /usr/src/drbd08
```

```
# make
```

```
# make install
```

# Configuração DRBD



- **Configuração do drbd através do drbdsetup.**
- **ou**
- **Configuração drbd via /etc/drbd.conf e drbdadm.**

**Anexo arquivo de configuração drbd.conf.**

# Configuração DRBD



**# modprobe drbd**

**Um cat em '/proc/drbd' deve resultar em algo muito semelhante ao seguinte:**

**version: 8.0 (api:82/proto:80)**

**SVN Revision: 2169 build by  
root@no1, 2006-06-30 10:51:39**

**0: cs:Unconfigured**

**1: cs:Unconfigured**

# Configuração DRBD



- Criando Meta-data.
- **# drbdadm create-md r0**
- Irá solicitar confirmação e o envio de informações para o site do drbd que faz a contagem de clusters no mundo.
- Levantando nós nas duas máquinas.
- **# drbdadm up r0**
- **# cat /proc/drbd**

# Configuração DRBD



- Tornando uma máquina primária.

```
# drbdadm primary r0
```

- Caso ocorra erro:

```
# drbdadm -- --overwrite-data-of-peer  
primary r0
```

- e acompanhe o progresso da sincronização em `'/proc/drbd'`

# Instalação OCFS2

**Via Gerenciador de pacotes (yum, apt-get)**

▪ **Pacotes necessários:**

**linux-image-2.6.x-server**

**libcomerr2, comerr-dev, uuid[-dev],  
libreadline5[-dev], libglib2.0-0, libglib2.0-  
dev**

**Instalar as ferramentas administrativas do  
OCFS2:**

**# apt-get install ocfs2-tools ocfs2console**

# Instalação OCFS2

**Via pacote TARBALL.**

**Site OCFS2:**

- **Descompacte o pacote ocfs2-x.x.x.tar.gz (no diretório /usr/src, de preferência).**
- **Siga as instruções do arquivo readme ou install.**

# Configuração OCFS2

- **OCFS2 possui apenas um arquivo de configuração**
- **`/etc/ocfs2/cluster.conf`**
- **onde são especificados os nós do cluster. Este arquivo deve ser o mesmo para todos os nós do cluster.**
- **Mostrar arquivo `cluster.conf`**

# Configuração OCFS2

## # /etc/init.d/o2cb load

- **Se não houver esse arquivo do sysv copiar do arquivo tarball descompactado.**

# Loading module "configfs": OK

# Creating directory '/config': OK

# Mounting configfs filesystem at /config: OK

# Loading module "ocfs2\_nodemanager": OK

# Loading module "ocfs2\_dlm": OK

# Loading module "ocfs2\_dlmfs": OK

# Mounting ocfs2\_dlmfs filesystem at /dlr

# Configuração OCFS2

- **Colocar o cluster online:**
- **# /etc/init.d/o2cb online darlan-cluster**
- **Iniciar durante o boot:**
- **# /etc/init.d/o2cb configure**

# OCFS + DRBD

## Formatando o FileSystem:

```
# mkfs.ocfs2 -b 4K -C 32K -N 2 -L  
darlan-cluster /dev/drbd0
```

## Montando o filesystem:

```
# mount -t ocfs2 /dev/drbd0 /dados
```

- <http://linux-ha.org/download/>
- Instalação via Tarball ou gerenciador de pacotes.

**Arquivos de configuração HA:**

**`/etc/ha.d`**

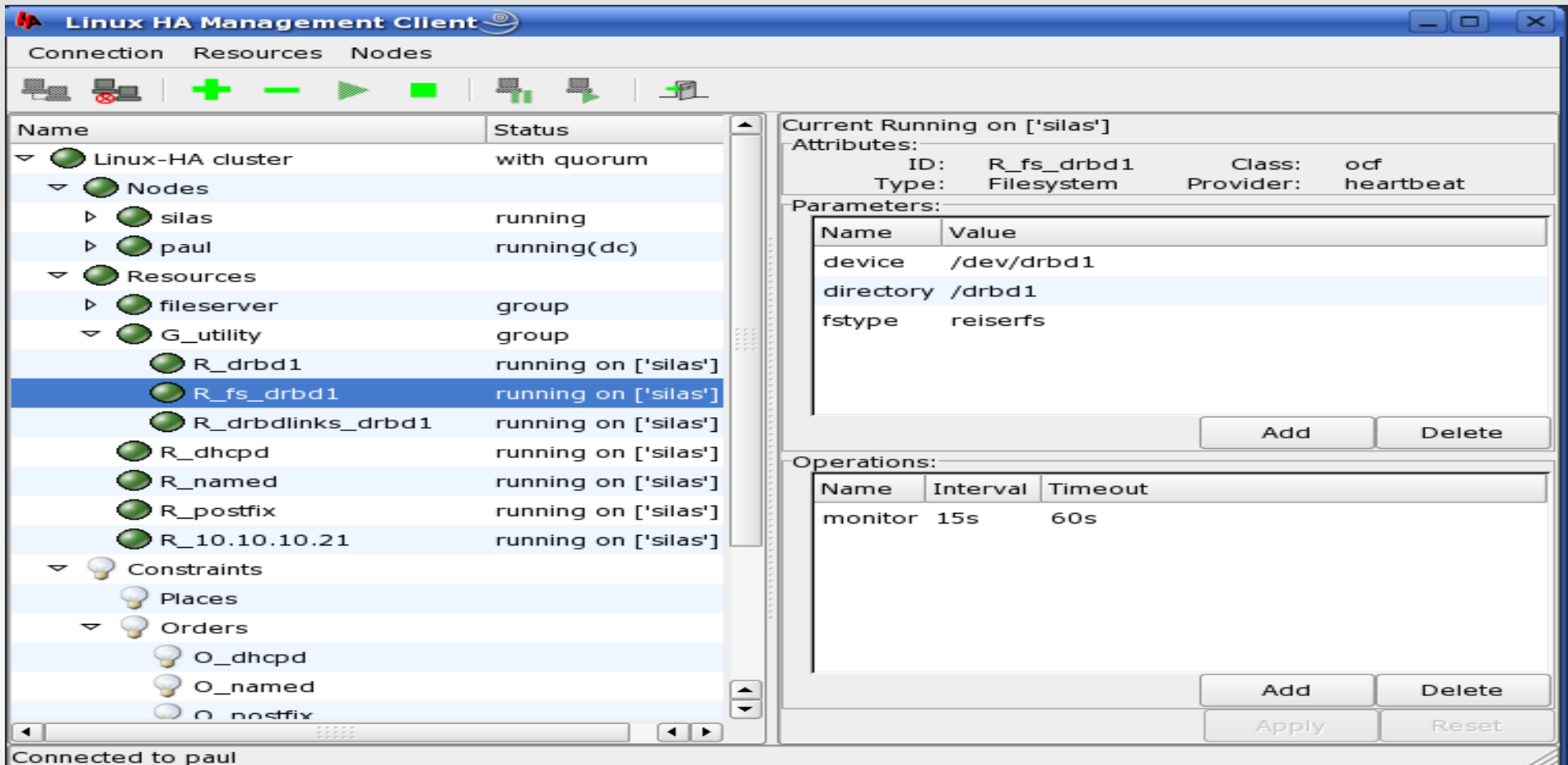
**Arquivos:**

**`ha.cf`, `haresources` e `authkeys`**

- **ha.cf = Configuração nós.**
- **haresources = Configuração serviços que Heartbeat irá monitorar.**
- **authkeys = Configuração autenticação de clusters com heartbeat.**

# Heartbeat

- **Configure it with `--enable-mgmt`.**



The screenshot shows the Linux HA Management Client interface. The left pane displays a tree view of the cluster configuration. The right pane shows the configuration for the selected resource, R\_fs\_drbd1, which is currently running on the 'silas' node.

**Linux HA Management Client**

Connection Resources Nodes

Name Status

- Linux-HA cluster with quorum
  - Nodes
    - silas running
    - paul running(dc)
  - Resources
    - fileserver group
    - G\_utility group
      - R\_drbd1 running on ['silas']
      - R\_fs\_drbd1 running on ['silas']**
      - R\_drbdlinks\_drbd1 running on ['silas']
      - R\_dhcpd running on ['silas']
      - R\_named running on ['silas']
      - R\_postfix running on ['silas']
      - R\_10.10.10.21 running on ['silas']
    - Constraints
      - Places
      - Orders
        - O\_dhcpd
        - O\_named
        - O\_postfix

Current Running on ['silas']

Attributes:

ID:	R_fs_drbd1	Class:	ocf
Type:	Filesystem	Provider:	heartbeat

Parameters:

Name	Value
device	/dev/drbd1
directory	/drbd1
fstype	reiserfs

Operations:

Name	Interval	Timeout
monitor	15s	60s

Connected to paul

# Considerações Finais

➤ **Solução ideal:**

**DRBD 7 || 8.**

**Ext3 || OCFS2.**

**HeartBeat.**

**Mon.**

# Referências

- <http://www.drbd.org>
- <http://oss.oracle.com/projects/ocfs2/>
- <http://www.linux-ha.org/>
- <http://guialivre.governoeletronico.gov.br/>
- <http://ha-mc.org/?q=node/15>

# Contato

**Darlan Segalin**

**darlanse@gmail.com**

**www.darlansegalin.net**